ISSN 0146-4116, Automatic Control and Computer Sciences, 2020, Vol. 54, No. 2, pp. 168–178. © Allerton Press, Inc., 2020.

Pedestrian Detection with EDGE Features of Color Image and HOG on Depth Images

**Xiong Zhang[a], *, Hong Shangguan[a], **, Aiping Ning[a], Anhong Wang[a],
Jiao Zhang[a], and Sichun Peng[a]**

[a] *School of Electronic Information Engineering, Taiyuan University of Science and Technology,
Taiyuan, Shanxi, 030024 P.R. China*

**e-mail: zx@tyust.edu.cn*

***e-mail: shangguan_hong@tyust.edu.cn*

Received September 2, 2019; revised October 30, 2019; accepted November 4, 2019

Abstract—The existing pedestrian detection algorithms are not robust in the case of noise, obstruction and illumination change. To solve the problem, we propose a pedestrian detection algorithm combining the edge features of color image with the features of Histogram of Oriented Gradient on Depth image (referred to as the HOD features). The algorithm describes overall structural features of pedestrians by using shearlet transform to extract their edge features from color images, and obtains local edge features of corresponding depth images by generating HOD features. The overall structural features and local edge features are combined to form new feature descriptors to train a SVM (support vector machine) classifier. Due to the full integration of the two types of features, the algorithm shows significant advantages in pedestrian detection in the case of interfering factors such as noise, obstruction, illumination, and similar colors. The experimental results show that the detection accuracy rate of this algorithm is 15% higher than that of other algorithms when the false-positive rate is 0.1.

Keywords: pedestrian detection, shearlet transform, edge feature, histogram of oriented depth, SVM classifier

DOI: 10.3103/S0146411620020108

1. INTRODUCTION

Pedestrian detection is a fundamental technique that uses computer vision to determine whether a pedestrian is present in an image or video sequence and to locate the pedestrian precisely [1, 2]. Due to that pedestrian movement is non-rigid, and that pedestrians may be in an ever-changing environment or dressed differently with various postures, there is much difficulty with pedestrian detection [3]. Much research has been conducted on how to accurately locate pedestrians [4–8]. At present, most methods realize pedestrian detection based on computer vision through feature extraction and machine learning. Haar wavelet coefficients [9] and Histograms of oriented gradients (HOG) [10] are the usually utilized features. Common machine learning methods include adaptive enhancement algorithm (Adaboost) [11], neural network [1, 5] and support vector machine (SVM) [12]. According to the different features adopted, pedestrian detection methods can be divided into three classes.

On one hand, single feature is extracted from color images. A most widely used pedestrian detection algorithm of this kind is HOG algorithm, which was firstly proposed by Dalal [10]. HOG is a dense descriptor for overlapping areas of images. The pedestrian features of this type of algorithm are constructed by calculating gradient direction histograms of local blocks that are allowed to overlap each other, thereby making the algorithm robust. Compared with Haar, HOG has the disadvantage of large amount of computation, and it is not suitable for applications where the number of traversing windows is large in rough level. Mu [13] combined the Local Binary Pattern (LBP) features with an SVM classifier to achieve pedestrian detection with very fast computation speed. However, these two algorithms are limited in their performance due to their use of single feature. Spinello and Arras [14] proposed Histogram of Oriented Gradient (HOD) feature on depth image inspired by HOG. HOD follows the same processing flow as HOG in depth image, and the obtained local depth change array can well describe the local appearance features.

On the other hand, multiple features are extracted from color images. The HOG features and the LBP features were combined for pedestrian detection [15]. A covariance matrix of various features (e.g., pixel coordinates, first and second derivative of grayscale, and gradient direction) was used to describe local features of pedestrians [16]. This kind of algorithms can provide a more detailed description of pedestrians through feature fusion, but are susceptible to illumination, obstruction, complex background and other factors because the algorithms rely on RGB color images for feature extraction, and consequently their robustness and accuracy in pedestrian detection is greatly limited.

In addition, Multi-features are extracted from depth images. Spinello et al. [14] first introduced depth images into pedestrian detection, and based on HOG they proposed the use of HOG on depth images to extract features (or simply referred to as Histogram of Oriented Depth features, shortly HOD features in their report), and performed weighted fusion of the HOG and HOD features to achieve pedestrian detection. Wang [17] combined the HOG, HOD and PDSS (pyramid depth self-similarities) features to achieve good pedestrian detection performance. Color images are rich in color and texture information, but have high requirements for illumination conditions. In contrast, depth images are independent of illumination conditions, so the two types of images are mutually supplementary and can be combined to enhance detection efficiency. There are two limitations when HOG is used to describe color images: On one hand, HOG only displays local textures but not structural features, so the gradients are biased in the case of noise or irregular edge curves; on the other hand, the gradients are approximated by finite difference, so the description of gradients is inaccurate.

In summary, image feature extraction methods play a key role in the performance of pedestrian detection algorithms. Firstly, although single color images are rich in color and texture data with good resolution and thereby have unique advantages in the detection process, they are susceptible to non-ideal conditions such as non-ideal illumination intensity, complex background, and obstruction, resulting in a decrease of the robustness and overall detection accuracy; although depth images are not affected by these conditions, they are subject to some disadvantages of their own such as low resolution and less texture information; the two types of images have mutually complementary advantages. Secondly, local image features have weak correlation with one another, and the disappearance of some local image features does not affect a stable detection and identification of other local features under obstruction, but they are more susceptible to noise.

In order to achieve pedestrian detection in a noisy environment, it is required that features extracted from color images be able to accurately describe the structural outline of pedestrians. Shearlet transform (ST) can fully capture the directions and other geometric features of pedestrians and accurately describe their structural properties through anisotropic, multi-scale transformation, thereby mitigating the adverse effects of noise, scaling and obstruction on pedestrian detection. Sheng et al. [18] applied ST to image edge analysis and detection, demonstrating that the transform can accurately capture geometric information of the edges and is very effective in the detection of edge positions and directions. Murugan et al. [19] used ST of translational invariance and neural network to achieve efficient and automatic detection of glioblastoma brain tumors.

This paper proposes a pedestrian detection algorithm in which edge features captured by ST are combined with HOD features to form new feature descriptors, which are used in conjunction with an SVM classifier for pedestrian detection. Edge feature descriptors extracted by ST not only enable full characterization of the structural outlines and other structural information of pedestrians, but can allow textures and other specific information to be described in detail. In addition, the outlines of targets in depth images are incomplete due to low image resolutions, but the edges of targets are clear and less susceptible to noise, so HOD features can be extracted to supplement color image features. Experiments show that feature descriptors incorporating both depth information and edge features captured by ST allow a more specific and accurate description of pedestrians, leading to a more robust detection performance in the case of complex environments including obstructions, and illumination change.

2. RELATED WORK

2.1. HOG

HOG is a dense descriptor for overlapping areas of images. It uses the distribution of local gradients to well describe the appearances and shapes of the local region with no need to know the exact positions of these gradients in the cells [20]. HOG features are also especially suitable for the description of pedestrian overall contour, and have high detection rate and low false alarm rate. A general HOG feature extraction and object detection chain diagram is shown in Fig. 1.

The specific steps of the algorithm are as follows:

a) *Preprocessing*. The input image is standardized in color space by Gamma correction method.

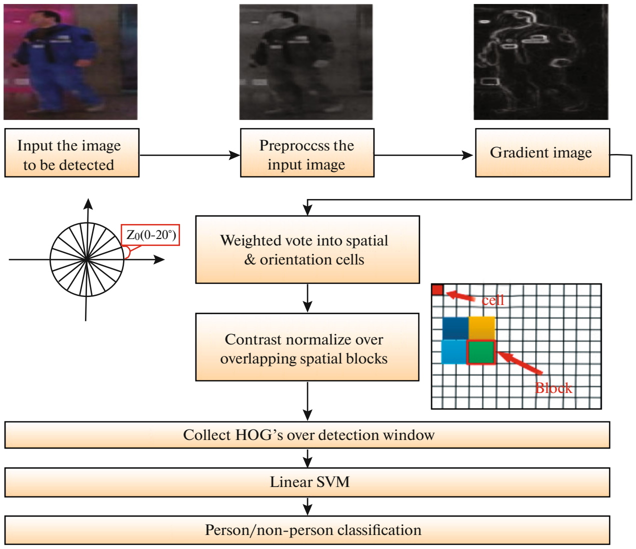
Input the image to be detected
Preproccss the input image
Gradient image
Z₀(0-20°)
Weighted vote into spatial & orientation cells
cell
Contrast normalize over overlapping spatial blocks
Block
Collect HOG's over detection window
Linear SVM
Person/non-person classification

Fig. 1. HOG feature extraction and object detection chain diagram.

b) *Gradients calculation*. Suppose the gradient at the image pixel coordinate (x, y) is denoted as $H(x, y)$, then the horizontal gradient is

$$G_x(x, y) = H(x+1, y) - H(x-1, y). \tag{1}$$

The vertical gradient is

$$G_x(x, y) = H(x, y+1) - H(x, y-1). \tag{2}$$

The Gradient amplitude is

$$G_x(x, y) = \left(G_x(x, y)^2 + G_y(x, y)^2\right)^{0.5}. \tag{3}$$

The gradient direction is

$$\alpha(x, y) = \tan^{-1}\left(G_x(x, y)/G_y(x, y)\right). \tag{4}$$

c) *Construction of HOG*. The whole image is divided into small connected regions called cells, and each pixel within one cell casts a weighted vote for a HOG channel based on the gradient amplitude of the pixel in that channel, and the weighted sum of projected amplitudes in each bin is calculated to establish a HOG for each cell. HOG can further weaken the effects of pedestrian motions and appearance on the detection. HOG comes in two types. The first type has histogram channels spreading over 0 to 180 degrees (unsigned gradient), and the second type over 0 to 360 degrees (signed gradient). HOG usually consists of nine channels with unsigned gradients. In this paper, the gradient orientation is divided into 9 bins over 0 to 180 degrees, with each cell having a size of 8 × 8 pixels.

d) *Normalization within the block*. Neighboring cells form a block, and there are adjacent blocks. Half of the areas overlap. The gradient direction histograms of four cells in block are connected in series and normalized using L2 norm paradigm to obtain the HOG features of block.

e) *Collection of HOG features*. By moving a sliding detection window on the image in specific steps, HOG features are collected to form a feature vector, which is subject to classification training in an SVM classifier to provide final results.

2.2. Shearlet

ST is a multi-scale geometric analysis method based on the affine system with composite dilations. It generates a nearly optimal multidimensional sparse representation by performing the affine transforms of scaling, shearing and translation on basic functions to produce shearlet functions with various features, and it is an extension of the wavelet transform framework in two-dimensional and higher-dimensional data spaces. ST is defined by formula (5):

$$SH_{\psi}f_{(a,s,t)} = \left\langle f, \psi_{a,s,t} \right\rangle,\ f \in L^2(R^2), \tag{5}$$

where f represents the original function and belongs to the two-dimensional square integrable space $L^2(R^2)$, and a, s, t represents the scale, direction, and transformation factor, respectively. Therefore, ST has the ability to capture multi-directional data, and therefore is able to effectively describe the detailed information in complex scenarios at different scales, directions and positions; $\psi_{a,s,t}$ is a ST function, which can be computed according to formula (6):

$$\psi_{a,s,t}(x) = \left|\det M_{a,s}\right|^{-1/2} \psi(M_{a,s}^{-1}x - st), \tag{6}$$

where $M_{a,s} = \begin{pmatrix} a & \sqrt{as} \\ 0 & a \end{pmatrix} = SA$; $A = \begin{pmatrix} a & 0 \\ 0 & \sqrt{a} \end{pmatrix}$ is an anisotropic dilation matrix representing a multi-scale partition; $S = \begin{pmatrix} 1 & \sqrt{s} \\ 0 & 1 \end{pmatrix}$ is a shear matrix for direction analysis. x represents an independent variable. In practical applications, usually let $A_0 = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}$, and $S_0 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. For any $\xi = (\xi_1, \xi_2) \in \hat{R}^2$, $\xi_1 \neq 0$, let ψ meet the following conditions: $\hat{\psi}(\xi_1,\xi_2) = \hat{\psi}_1(\xi_1)\hat{\psi}_2(\xi_2/\xi_1)$, where $\hat{\psi}$ represents the Fourier transform of ψ; and $\hat{\psi}_1 \in C^{\infty}(R)$; $\hat{\psi}_1 \subset [-1/2,-1/16] \bigcup [1/16,1/2]$; $\hat{\psi}_2$ is a Bump function, $\hat{\psi}_2 \subset [-1,1]$; $\hat{\psi} \in C^{\infty}(R)$ is continuous and compactly supported, with $\hat{\psi} \subset [-1/2, 1/2]^2$. In the frequency domain:

$$\hat{\psi}_{a,s,t}\left(\xi_1,\xi_2\right) = a^{3/4} e^{-2\pi i \xi t} \hat{\psi}\hat{\psi}_1\left(a\xi_1\right)\hat{\psi}_2\left(a^{-1/2}\left(\xi_2/\xi_1 - s\right)\right). \tag{7}$$

It can be seen that shearlets are more pronounced in the frequency domain, and each ST function $\psi_{a,s,t}$ is supported on a pair of trapezoids along the slope of s that are symmetric with respect to the origin. Therefore, each ST is a function set with the parameters of scale a, direction s, and position t.

3. METHOD

Traditional algorithms largely ignore structural properties of pedestrians when extracting color image features, while structural properties have very obvious advantages in the case of noise or obstruction. Color images are susceptible to non-ideal conditions, so extracting features from color images alone will decrease the robustness. Therefore, this paper proposes a pedestrian detection algorithm that combines color image edges and HOD. The algorithm extracts edge features of color images by ST so as to overcome the shortcomings of only focusing on local textures and ignoring the overall structure information when describing pedestrians. At the same time, the algorithm uses depth images to provide supplementary advantages for color images. By combining the edge features extracted from color images with the HOD features extracted from depth images, it is possible to achieve a more accurate description of pedestrians so as to enhance detection performance.

Each color image and its corresponding depth image are treated as one image pair, with the pair containing any pedestrian being a positive sample and the pair without any pedestrian being a negative sample. During the detection process, the color image is subject to ST to extract the edge features so as to obtain the overall structural outline information of the pedestrians, while the depth image is subject to HOD feature extraction to obtain local features of the pedestrians, with the two types of features combined to form new descriptors to input in an SVM classifier where they are used to train the classifier. Next, the images to be tested are input in the trained classifier where pedestrian positions in the images are subject to detection test based on a detection algorithm whose flowchart is shown in Fig. 2.

3.1. Edge Features Captured by ST

Edges refer to pixels that undergo large pixel value changes, and edge feature extraction is targeted at finding the edges of detection targets in images. ST not only scales the image, but also performs geometric transformations such as rotation and shear transformation, so shearlets have high directional sensitivity, and the number of directions will gradually double with a refinement of the scales. Moreover, ST can describe geometric characteristics of the singularity (i.e., singular points, straight lines, curves) of two-dimensional images, because the decay speed of $SH_{\psi}\ f(a,s,t)$ not only describes the positions, but also

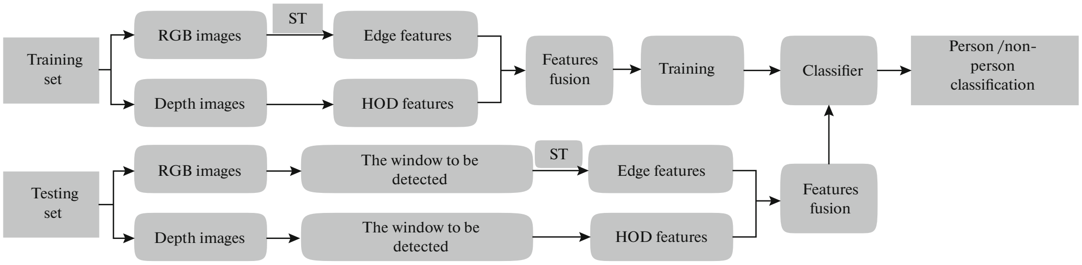

Fig. 2. Algorithm flowchart.

reflects the directions of the singularity. Therefore, ST has the ability to capture the directions of multidimensional data and is the best tool to describe image edges. Edge features captured by ST approximate the original outlines of the targets in nearly a linear optimal manner. Two types of image are included in this study – color images and depth images. However, depth images have a low resolution with incomplete outlines of targets, so they are unable to very accurately describe pedestrian targets, while color images are so rich in texture features with clear pedestrian outlines that they can provide a better description of pedestrian targets. For this reason, edge features are extracted only from color images in this study. The following theorem is the basis of shearlet-based edge detection, illustrating the approximation ability of ST [18].

Theorem: If $t_1 + t_2 = 1$ (assuming that the circle represents the curve) and $s = t_2/t_1$, $t_1 \neq 0$, one has the following: when $a \to 0$, $SH_\psi f(a,s,t) \to a^{3/4}$; in other cases, when $a \to 0$, $SH_\psi f(a,s,t) \to 0$. In order to achieve discretization of ST, let scale parameter $a_j = 2^j$, shear parameter $s_{j,k} = k \times 2^{j/2}$ and translation parameter $t = m$, and accordingly one has formula (8)

$$\psi_{j,k,m}(x) = 2^{\frac{3j}{4}} \psi(B_1^k A_1^j x - m), \tag{8}$$

where $j \geq 0$, $-2^{j/2} \leq k \leq 2^{j/2}$, $m \in R^2$. Shearlet-based edge detection steps:

(1) Performing ST on the image $f \in L^2(R^2)$.

(2) Obtaining $SH_\psi^{[0]} f_{(j,k,m)}$ and $SH_\psi^{[1]} f_{(j,k,m)}$ which are associated with the horizontal cone and the vertical cone, respectively.

(3) Predicting the edge e_j of the j-th image layer according to the following formula:

$$e_j f[m] = \sqrt{\sum_k (SH_\psi^{[0]} f_{(j,k,m)})^2 + \sum_k (SH_\psi^{[1]} f_{(j,k,m)})^2}. \tag{9}$$

3.2. HOD

A depth image is also a distance image in the sense that the distances from the image collector to the points in a scene are used as image pixels, and such type of distance image directly reflects collective shapes of the visible surfaces in the scene. Histogram of Oriented Gradient on depth image (HOD) is a feature on the depth image obtained by following the same steps as HOG on color image. The steps to extract the HOD features are as following: dividing the window into cells, calculating the descriptor of each cell, compiling oriented gradients into a one-dimensional histogram. Every four cells are grouped to form a block, which is subject to normalization by the L2-Hys method in order to ensure good robustness to depth noise. The rationality behind the above processes is straightforward, that is, the data set of local depth variation can be used to give a good description of the local 3D shapes and appearance. The resulting HOD feature vector is used to train a linear SVM classifier.

However, real distances are unevenly coded in original depth images, with the following relationship between the original depth v and the distance d (in meters):

$$d = \frac{8BF_x}{V_{\max} - v}. \tag{10}$$

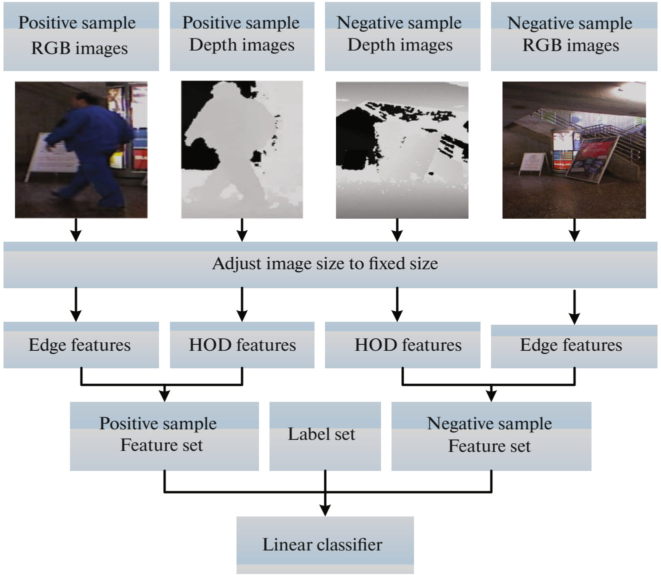
Positive sample RGB images
Positive sample Depth images
Negative sample Depth images
Negative sample RGB images
Adjust image size to fixed size
Edge features
HOD features
HOD features
Edge features
Positive sample Feature set
Label set
Negative sample Feature set
Linear classifier

Fig. 3. Flowchart of the training process.

In formula (10), B denotes the distance from the infrared emitter to the infrared camera, F_x is the horizontal focal length of the infrared camera, and d can be ignored if negative. The above formula is a hyperbolic relationship, which implies that when the depth is beyond a certain range, the depth resolution will become quite low and therefore it is necessary to preprocess the original depth image to well separate the foreground and background, namely multiplying the distance in meters by $M/D\max$ where $M = 100$ representing a constant gain, and $D\max = 20$ representing the maximum distance in meters.

3.3. Training

The training flowchart is shown in Fig. 3, and the detailed steps are as follows: (a) entering the positive and negative sample pairs to be trained, and setting all the images to a fixed size; (b) labeling the positive samples as 1 and the negative samples as –1; (c) extracting edge features of the color images in the positive and negative samples by ST; (d) extracting HOD features of the depth images in the positive and negative samples; (e) concatenating the features extracted in each image pair as a new feature descriptor; (f) inputting the feature descriptors of all positive and negative samples and their label features into a linear classifier for training; (g) obtaining classification vectors of the classifier.

3.4. Test

The flowchart of pedestrian detection test is shown in Fig. 4 with the steps detailed as below: (a) entering the pair of pictures to be tested; (b) scaling the test image pair until the image size is smaller than a fixed window size; (c) scrolling a fixed-size window throughout each image at various scales and extracting features from the image section in each window; (d) concatenating the features of each image pair in a fixed-size window to obtain a fused feature descriptor; (e) inputting the fused feature descriptor into the classifier to obtain a corresponding classification value; (f) comparing the obtained value with the threshold to determine whether the window contains any pedestrian.

4. RESULTS AND DISCUSSION

The feasibility and superiority of the proposed method were verified by conducting experiments using the data set, which is derived from the database provided by Luciano and Luber [21], where some color images and their corresponding depth images are selected as the training set while the remaining images are used as the test set. We use the classical linear SVM classifier because of its excellent performance and wide application. The experimental environment is a computer with I5Intel Xeon, 3.6 GHz CPU and 32GB RAM equipped with a MATLAB 2014b platform as the algorithm development environment. The

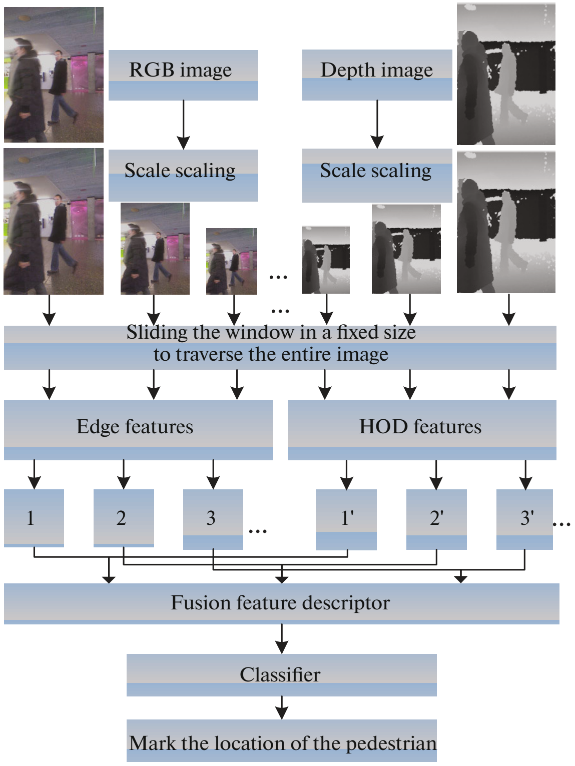
RGB image
Depth image
Scale scaling
Scale scaling
...
...
Sliding the window in a fixed size
to traverse the entire image
Edge features
HOD features
1
2
3
...
1'
2'
3'
...
Fusion feature descriptor
Classifier
Mark the location of the pedestrian

Fig. 4. Flowchart of the test.

comparison algorithms adopted by this study are three popular tracking algorithms, namely HOG, HOG + HOD, and HOG + PDSS. Relevant parameters of our algorithm are as follows: the number of positive sample pairs is 2920 and negative sample pairs is 11589, the cell size is set to 8 × 8 pixels for HOG extraction, each block consists of 4 cells, the fixed size of the sliding window is 128 × 64 pixels, the number of bins is 9, and the orientation ranges from 0 to 180 degrees.

The following test indicators are involved in the detection performance evaluation of the algorithms: TP (true-positive) samples: the positive samples that are correctly classified, the larger the quantity, the better the performance; TN (true-negative) samples: the negative samples that are correctly classified, the larger the quantity, the better the performance; FP (false-positive, false positive) samples: the number of negative samples that are misclassified, the smaller the quantity, the better the performance; FN (false-negative) samples: the number of positive samples that are misclassified, the smaller the quantity, the better the performance; oks: the samples that are correctly classified, the larger the quantity, the better the performance; kos: the samples that are misclassified, the smaller the quantity, the better the performance.

Table 1 contains the basic test indicator data of various algorithms – the numbers of oks, kos, TP, TN, FP, and FN samples. For a specific indicator, the algorithm name underlined and highlighted in bold red refers to the optimal algorithm in terms of this indicator, while the algorithm name in italic, bold blue with a wavy line refers to the suboptimal algorithm. In the second column of oks, the highest data is at the second row corresponding to the algorithm of this study, indicating that this algorithm accurately detects the largest number of samples and its detection performance is the best. In the third column of kos, the smallest data is associated with the algorithm of this study, indicating that this algorithm misclassifies the smallest number of samples. The columns of TP and TN represent the numbers of correctly detected positive and negative samples, respectively, with a larger number indicating a better detection performance, while the columns of FP and FN represent the numbers of incorrectly classified negative and positive samples, with a smaller number indicating a better detection performance. For these four indicators, the results of our algorithm—as listed in the second row—are not the best, but are very close to those of the optimal algorithms in this case. Moreover, the indicator results of our algorithm are good on the whole, better than those of other algorithms.

Based on the basic variables, some important evaluation indicators are derived: FPR (false-positive rate)—the probability of misclassifying negative samples; TPR (true-positive rate)—the probability of correctly classifying positive samples; accuracy—the probability of correctly classifying positive and negative

Table 1. Data of the basic test indicator data for different algorithms

	oks	kos	FP	FN	TP	TN
This work	**17518**	**3067**	***24***	***3043***	***5953***	***11565***
HOG	15881	4704	**6**	4698	4298	**11583**
HOG + HOD	***17462***	***3123***	585	**2538**	**6458**	11004
HOG + PDSS	15462	5123	244	4879	4117	11345

Table 2. Key indicators of various algorithms

	FPR	TPR	Accuracy	FRR	FAR
This work	***0.21%***	***66.17%***	**85.1%**	***33.82%***	***99.79%***
HOG	**0.05%**	47.78%	77.15%	52.22%	**99.95%**
HOG + HOD	5.05%	**71.79%**	***84.83%***	**28.21%**	94.95%
HOG + PDSS	2.1%	45.76%	75.11%	54.24%	97.89%

samples; FRR (false reject rate)—the probability of misclassifying positive samples; FAR (false accept rate)—the probability of correctly classifying negative samples.

In Table 2, the data that are underlined and in red bold are the optimal values of the indicators, while the data that are underlined, in italic with a bold wavy line are the suboptimal values. As shown from Table 2, our algorithm has suboptimal performance in the four indicators of FPR (second column), TPR (third column), FRR (fifth column), and FAR (sixth column) as listed at the second row, but when pooled together, the four indicators have better results on the whole than those of other algorithms. The fourth column of "accuracy" represents the overall detection accuracy—an indicator for which our algorithm gives the highest value, indicating that the algorithm has the best performance.

In order to verify the superiority of the algorithm in a more straightforward manner, the ROC (receiver operating characteristic) curve is used as the evaluation index, as shown in Fig. 7. The X-axis represents FPR and the Y-axis TPR. The area under the ROC curve (AUC) is a metric, which takes a larger value when the ROC curve is shifted more toward the upper left corner—a scenario suggesting that given the same FPR, the higher the TPR the better the performance.

Figure 5 depicts the ROC curves of our algorithm and three other popular algorithms. The results show that our algorithm leads to the greatest AUC and its ROC curve is closest to the upper left corner than other algorithms. When FPR is 0.1, our algorithm has the highest detection accuracy of all the four algorithms, which is 15% higher than that of the suboptimal algorithm, that is, our algorithm has the best detection performance.

The detection performances of the four algorithms of this work, HOG+HOD, HOG+PDSS, and HOG under different challenging conditions are listed below respectively, in order of left to right. The region inside a green rectangle frames refers to a pedestrian position detected by the algorithms, while the region inside a red dotted oval frame is the image region of special detection interest.

Figure 6 shows the results of four algorithms in the presence of slight obstructions, improper scales, and posture changes. In Fig. 6a, what are encompassed by the oval frames are pedestrians who are too close to the camera and thereby have too large image scales. For these pedestrians, their positions fail to be detected by HOG+PDSS (Fig. 6c) but are detected by the other three algorithms. In Fig. 6b, what are encompassed by the oval frames are pedestrians who are slightly obstructed for the camera, and it is evident that HOG+PDSS is only able to detect half parts of the pedestrians while failing to identify the parts under slight obstruction and thereby unable to locate the whole positions of the pedestrians. HOG, HOG+HOD, and our algorithm can accurately detect the positions of the pedestrians, and in particular our algorithm gives the most accurate positions. In Fig. 6c, the pedestrians in the oval frame have a large stride different from an average pedestrian stride, and the whole bodies—except the faces—are not accurately located by HOG+PDSS, while the other three algorithms correctly locate the full positions of the pedestrians. In short, our algorithm performs well in the case of slight obstruction, over-scaling and posture change.

Figure 7 shows the detection results of four algorithms in the case of strong obstruction and attitude change. In Fig. 7a, what is encompassed by the red oval frames is a pedestrian under strong obstruction whose half body is masked by the staircase. Of the four algorithms, only our algorithm can locate the pedestrian while the other three fail to. In Fig. 7b, the four algorithms all locate the position of the leftmost pedestrian inside the

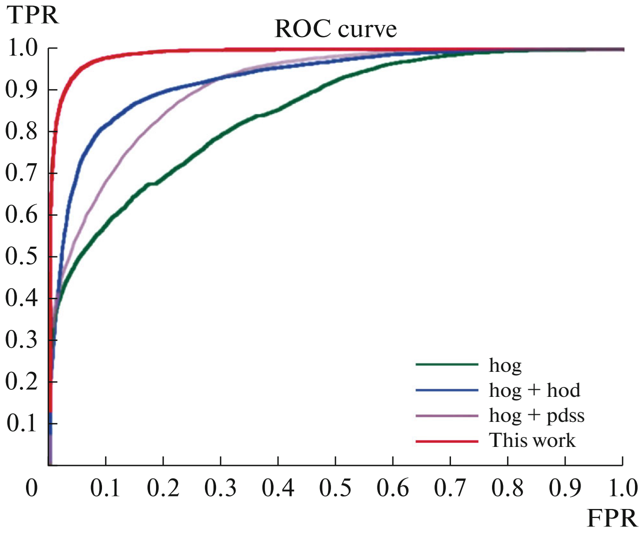

Fig. 5. Comparison of ROC curve between our algorithm and the other three algorithms.

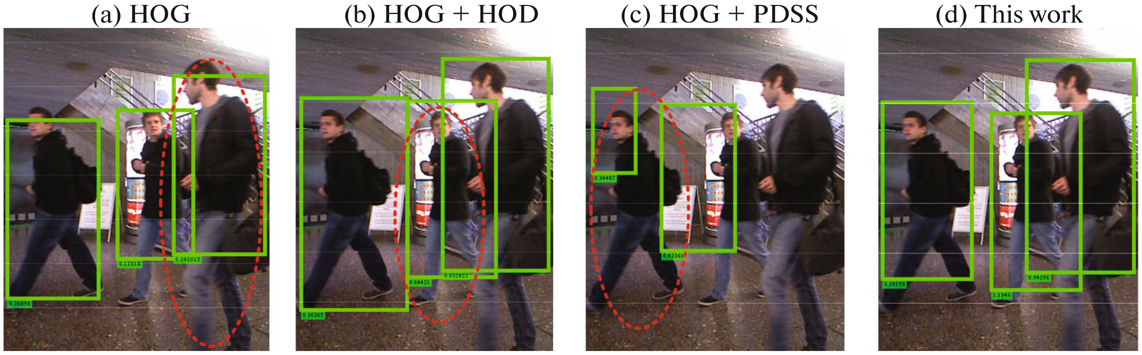

Fig. 6. Detection results of four algorithms in the case of slight obstruction, improper scale, and posture change.

oval frame, but due to the special posture of the pedestrian, the detected position is not fully accurate. In Fig. 7c, the pedestrian in the middle of the oval frame is obstructed by a bag in the front, and all the four algorithms can find the accurate position of the pedestrian, while HOG+PDSS fails to provide a fully accurate position. In short, our algorithm performs well in the case of strong obstruction and attitude change.

Figure 8 shows the detection results of four algorithms in the presence of strong illumination and severe obstruction. In Fig. 8a, two pedestrians under severe obstruction are encompassed in the oval frame, which are so difficult to visually observe due to nearly 90% of their bodies being obstructed that the two pedestrians are misclassified as one person by all the four algorithms. In particular, the frame position of HOG is not very accurate. In Fig. 8c, the pedestrian in the red rectangular frame fails to be located by HOG+PDSS as the pedestrian is in a backlight area, while the other three algorithms identify the pedestrian position accurately. In Fig. 8d, the pedestrian encompassed by the red oval frame is in a shadow area with strong illumination background, whose position is identified by all the four algorithms. In short, our algorithm performs well in the case of strong illumination.

Figure 9 shows the detection results of four algorithms in the presence of local illumination, strong obstruction, and improper scaling. In Fig. 9a, two pedestrians of large scales are encompassed by the red oval frame, whose positions are identified by the four algorithms. However, due to their too short distances to the camera and too large scales, the detected positions are not very accurate. In Fig. 9b, the pedestrian in the red oval frame is almost entirely obstructed, so none of the four algorithms can locate the pedestrian.

In Fig. 9c, a pedestrian under local illumination is encompassed by the red oval frame, and the illumination is so strong that except HOG+PDSS, the other three all find the position while failing to explicitly identify the pedestrian. In short, our algorithm performs well in the case of illumination change.

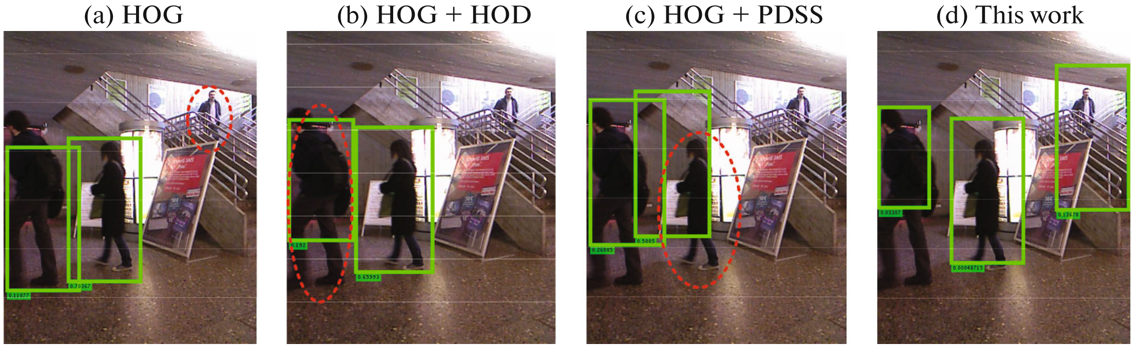

Fig. 7. Test results of four algorithms in the case of strong obstruction and different postures.

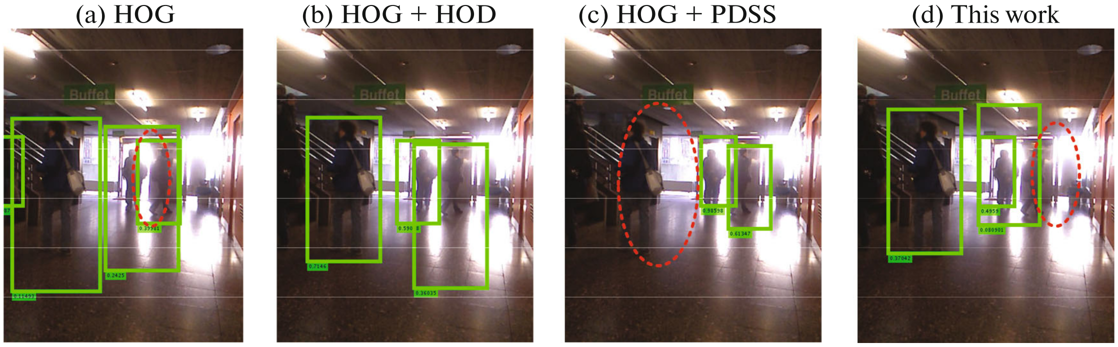

Fig. 8. Detection performances of four algorithms in the presence of strong illumination and severe obstruction.

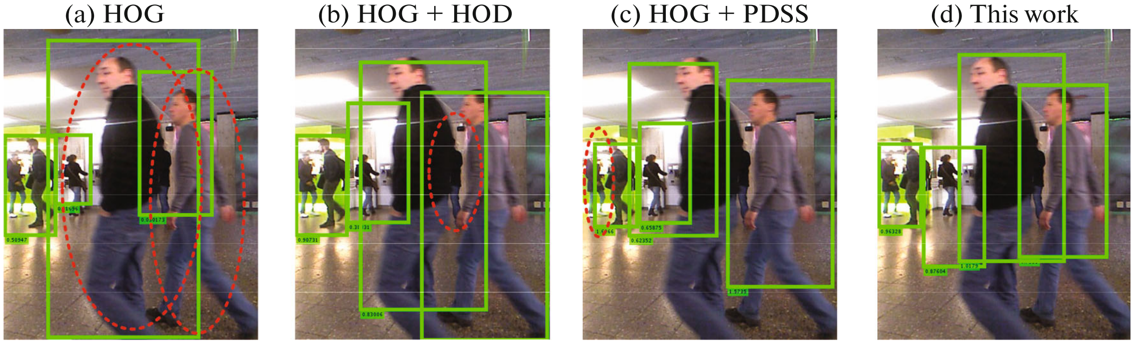

Fig. 9. Detection performances of four algorithms in the presence of local illumination, strong obstruction, and improper scaling.

5. CONCLUSIONS

This paper proposes a pedestrian detection algorithm that combines color image edge information and Histogram of Oriented Gradient on depth image. In order to overcome the shortcomings that most algorithms fail to provide accurate feature description due to their focusing on the local features and ignoring the overall structure. Our algorithm provides more accurate description of the overall structural features of pedestrians by extracting edge features from color images through shearlet transform, and obtains local gradient and edge features in depth images by extracting Histogram of Oriented Gradient on depth images, followed by combining the two types of features to form new feature descriptors so as to achieve more accurate description of pedestrians. Experiments show that our algorithm can give a 15% higher detection accuracy rate compared with other algorithms. Moreover, our algorithm is superior to other algorithms in the case of challenging conditions such as obstruction, illumination and similar colors.

Although the detection performance of our algorithm is satisfactory, it is still subject to low detection speed. There are two reasons for this defect. First of all, this algorithm uses a large number of information sources; secondly, the features are simply concatenated to form new descriptors without consideration of the repetition of information between features. Such an issue may be solved by performing feature compression to reduce the data volume, thereby increasing the detection speed.

FUNDING

This work has been supported in part by Scientific and Technological Innovation Team of Shanxi Province (no. 201705D131025) and Collaborative Innovation Center of Internet+3D Printing in Shanxi Province (201708).

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

1. Zhang, L., Lin, L., Liang, X., et al., Is faster R-CNN doing well for pedestrian detection?, *European Conference on Computer Vision,* Cham, 2016, pp. 443–457.
2. Combs, T.S., Sandt, L.S., Clamann, M.P., et al., Automated vehicles and pedestrian safety: Exploring the promise and limits of pedestrian detection, *Am. J. Prev. Med.,* 2019, vol. 56, no. 1, pp. 1–7.
3. Zhang, S., Benenson, R., Omran, M., et al., How far are we from solving pedestrian detection?, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2016, pp. 1259–1267.
4. Zhang, S., Benenson, R., and Schiele, B., Filtered channel features for pedestrian detection, *CVPR,* 2015, vol. 1, no. 2, p. 4.
5. Li, J., Liang, X., Shen, S.M., et al., Scale-aware fast R-CNN for pedestrian detection, *IEEE Trans. Multimedia,* 2017, vol. 20, no. 4, pp. 985–996.
6. Cai, Z., Saberian, M., and Vasconcelos, N., Learning complexity-aware cascades for deep pedestrian detection, *Proceedings of the IEEE International Conference on Computer Vision,* 2015, pp. 3361–3369.
7. Mao, J., Xiao, T., Jiang, Y., et al., What can help pedestrian detection?, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2017, pp. 3127–3136.
8. Paisitkriangkrai, S., Shen, C., and Hengel, A., Pedestrian detection with spatially pooled features and structured ensemble learning, *IEEE Trans. Pattern Anal. Mach. Intell.,* 2015, vol. 38, no. 6, pp. 1243–1257.
9. Li, J., Gong, W., Li, W., et al., Robust pedestrian detection in thermal infrared imagery using the wavelet transform, *Infrared Phys. Technol.,* 2010, vol. 53, no. 4, pp. 267–273.
10. Dalal, N. and Triggs, B., Histograms of oriented gradients for human detection, *International Conference on Computer Vision and Pattern Recognition (CVPR'05),* 2005, vol. 1, pp. 886–893.
11. Freund, Y. and Schapire, R.E., A decision-theoretic generalization of on-line learning and an application to boosting, *J. Comput. Syst. Sci.,* 1997, vol. 55, no. 1, pp. 119–139.
12. Cheng, H., Zheng, N., and Qin, J., Pedestrian detection using sparse Gabor filter and support vector machine, *IEEE Proceedings. Intelligent Vehicles Symposium,* 2005, pp. 583–587.
13. Mu, Y., Yan, S., Liu, Y., et al., Discriminative local binary patterns for human detection in personal album, *2008 IEEE Conference on Computer Vision and Pattern Recognition,* 2008, pp. 1–8.
14. Spinello, L. and Arras, K.O., People detection in RGB-D data, *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems,* 2011, pp. 3838–3843.
15. Wang, X., Han, T.X., and Yan, S., An HOG-LBP human detector with partial occlusion handling, *2009 IEEE 12th international conference on computer vision*, 2009, pp. 32–39.
16. Tuzel, O., Porikli, F., and Meer, P., Pedestrian detection via classification on Riemannian manifolds, *IEEE Trans. Pattern Anal. Mach. Intell.,* 2008, vol. 30, no. 10, pp. 1713–1727.
17. Wang, N., Gong, X., and Liu, J., A new depth descriptor for pedestrian detection in RGB-D images, *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012),* 2012, pp. 3688–3691.
18. Yi, S., Labate, D., Easley, G.R., et al., A shearlet approach to edge analysis and detection, *IEEE Trans. Image Process.,* 2009, vol. 18, no. 5, pp. 929–941.
19. Arunachalam, M. and Royappan Savarimuthu, S., An efficient and automatic glioblastoma brain tumor detection using shift invariant shearlet transform and neural networks, *Int. J. Imaging Syst. Technol.*, 2017, vol. 27, no. 3, pp. 216–226.
20. Soni, R., Kumar, B., and Chand, S., Text detection and localization in natural scene images based on text awareness score, *Appl. Intell.,* 2018, vol. 49, pp. 1376–1405.
21. Spinello, L., Luber, M., and Arras, K.O., Tracking people in 3D using a bottom-up top-down people detector, *International Conference on Robotics and Automation (ICRA),* Shanghai, 2011, pp. 1304–1310.